Generative Transformer Chatbots for Mental Health Support: A Study on Depression and Anxiety

Jordan J. Bird
Department of Computer Science, Nottingham Trent University
Nottingham, United Kingdom
jordan.bird@ntu.ac.uk

Ahmad Lotfi
Department of Computer Science, Nottingham Trent University
Nottingham, United Kingdom
ahmad.lotfi@ntu.ac.uk

ABSTRACT

Mental health is a critical issue worldwide and effective treatments are available. However, incidence of social stigma prevents many from seeking the support they need. Given the rapid developments in the field of large-language models, this study explores the potential of chatbots to support people experiencing depression and anxiety. The focus of this research is on the engineering aspect of building chatbots, and through topology optimisation find an effective hyperparameter set that can predict tokens with 88.65% accuracy and with a performance of 96.49% and 97.88% regarding the correct token appearing in the top 5 and 10 predictions. Examples of how optimised chatbots can effectively answer questions surrounding mental health are provided, generalising information from verified online sources. The results of this study demonstrate the potential of chatbots to provide accessible and anonymous support to individuals who may otherwise be deterred by the stigma associated with seeking help for mental health issues. However, the limitations and challenges of using chatbots for mental health support must also be acknowledged, and future work is suggested to fully understand the potential and limitations of chatbots and to ensure that they are developed and deployed ethically and responsibly.

CCS CONCEPTS

• Information systems → Information retrieval; • Theory of computation → Design and analysis of algorithms; • Human-centered computing → Interactive systems and tools.

KEYWORDS

Chatbots, Natural Language Processing, Transformers, Mental Health

1 INTRODUCTION

Mental health is a critical issue that affects millions of people around the world. According to the World Health Organisation (WHO), an estimated 5% of all adults suffer from depression [WHO, 2021]. The WHO also note that, although effective treatment is available, 75% of those categorised as low- and middle-income do not receive treatment. Indeed, awareness and acceptance of poor mental health have steadily improved [Frank and Glied, 2006, Jones and Wessely, 2005], but there is still a significant stigma about the need for professional help [Sickel et al., 2014]. Mental health stigma can act as a barrier for people experiencing depression, anxiety, or other mental health challenges, preventing them from accessing the support they need. The prevalence of mental health stigma has led many people to view online alternatives favourably over physical human interaction [Hanley and Wyatt, 2021].

This knowledge leads to the concept of the online chatbot. In recent years, advances in Natural Language Processing (NLP) have led to the development of chatbots as a tool for promoting mental well-being. Chatbots are computer programs that can simulate a natural conversation, providing support through textual input and output. Given their accessibility and anonymity, they have the potential to help alleviate the stigma associated with seeking help for mental health issues [Abd-Alrazaq et al., 2019].

This paper focuses on the engineering aspect of chatbots for mental health support, with a specific focus on answering questions about depression and anxiety. The study will explore hyperparameter space to build chatbots based on attention mechanisms and transformers, which are large language models. These models have shown great success in various natural language processing tasks and have the potential to provide effective and engaging support to individuals experiencing mental health challenges. Furthermore, the paper will present examples of interactions with optimised chatbots to demonstrate their effectiveness and usability. The main goal of this work is to contribute to ongoing research in the field of mental health and technology by exploring the potential of chatbots to provide accessible and effective support for people experiencing depression and anxiety.

The remaining parts of this paper are organised as follows; background and related work is presented in Section 2 followed by the proposed method in Section 3. The results and observations are presented in Section 4. Section 5 presents the conclusion and future work.

2 BACKGROUND AND RELATED WORK

Chatbots are Human-Computer Interaction (HCI) models that allow users to converse with machines through natural language [Bansal...
and Khan, 2018]. Most often in the modern literature, chatbots make use of artificial intelligence and machine learning to process an input and produce a response in the form of text [Suhaili et al., 2021] and have grown rapidly more prominent in research since the year 2015.

A recent scoping review of chatbots in mental health revealed several pieces of interesting information within the field [Abd-Alrazaq et al., 2021]. Namely, the majority of chatbots focus on support for depression and autism, and controlled the conversation for therapy, training, and screening. The approach in this work is that of question-answering; that is, the goal of the model is to generalise online resources to provide answers that people may have about the included categories.

Bhagchandani and Nayak proposed the combination of two natural language processing models for a mental health chatbot framework [Bhagchandani and Nayak, 2022]. In this study, the authors first perform text classification using sentiment analysis to discern whether the user should be directed to a chatbot for a generic chat or another for therapy-based conversation. A similar approach was proposed in CareBot [Crasto et al., 2021], where conversational data was used along with the PHQ-9 and WHO-5 screening questionnaires to train a chatbot using a multimodal approach. The study recorded lower perplexity values for transformers compared to recurrent methods, but experimental observations revealed that 63% of the participants preferred the response generated by the Transformer over 22% for Long Short Term Memory (LSTM) networks and 15% for the Recurrent Neural Network (RNN).

In 2021, Deshpande and Warren proposed an additional module for a mental health chatbot which could detect users at risk of self-harm [Deshpande and Warren, 2021]; In their study, text classification experiments noted that the Bidirectional Encoder Representations from Transformers (BERT) could achieve 97% accuracy in recognising the risk within scraped Reddit data that were not part of the training dataset. BERT representations were also applied in a recent work, which found that it was a promising approach compared to classical approaches for the detection of mental health status from Reddit posts [Jiang et al., 2020]. Alongside the use of attention, several other methods have also been proposed to improve chatbots. These include data augmentation by paraphrasing [Bird et al., 2021, Joglekar, 2022], transfer learning [Prakash et al., 2020, Syed et al., 2021], reinforcement learning [Cuayáhuittl et al., 2019, Liu et al., 2020], and ensemble learning [Almansor et al., 2021, Bali et al., 2019].

Transformers are a new type of neural network that have recently seen a rapid rise in popularity, achieving state of the art performance in natural language processing, image captioning, image synthesis, classification, and audio processing [Lin et al., 2022]. Most relevant to this study are the studies exploring how transformer models achieve the current best performance metrics for the synthesis of text and answering of questions [Devlin and Chang, 2018, Lukovnikov et al., 2019, Radford et al., 2019, Shao et al., 2019]. According to the original paper [Vaswani et al., 2017], the attention values are calculated as the scaled dot product; Weights are calculated for each token within the input text as follows:

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V
\]

where \(Q\) is the query token, an embedded representation of a word within a sequence. \(K\) represents keys, vectors of the sequence of tokens presented to the model, and \(V\) are values that are calculated when querying keys. In this study, \(Q, K,\) and \(V\) are from the same data source, and therefore the operation is described as self-attention. Each block also contains several attention heads, and thus the approach that this study implements is known as multi-headed self-attention (\(MH\)). This is simply calculated via the concatenation of \(h_i\) attention heads as follows:

\[
MH(Q, K, V) = \text{Concatenate}(h_1, ..., h_k) W^Q
\]

The application of multi-headed attention has shown a significant improvement in ability compared to the conventional approach.

It is suggested that a shallower, wider model is more stable during the training process. Fig. 1 shows a diagram of how the model uses embeddings as input and output, with a tokenizer used to transform both strings into encodings and vice versa.

3 METHOD

Within this section, the proposed methodology will be discussed, followed by work on optimisation of chatbots to answer mental health questions. The general approach of this work can be observed in Fig. 2; this section details each step of this process.

Initially, data from various sources were collected to form a large dataset. No single modern dataset is viable for large neural language models given their data requirements for effective generalisation [Saggin et al., 2022]. Due to this, data from CounselChat\(^1\), the Brain & Behaviour Research Foundation\(^2\), the NHS\(^3\), Wellness in Mind\(^4\) and White Swan Foundation\(^6\) were selected. Questions and answers are extracted, and questions are manually generated dependent on the information available, e.g. for the NHS definition of depression, questions such as “what is depression?” are imputed.

For preprocessing, all texts were converted to lowercase, and punctuation was removed in order to reduce the learning of irrelevant tokens. For example, the tokens “Hello”, “hello”, “Hello!” and “hello?” would all be treated as separate learnable tokens prior to this step. Then the vocabulary was limited to the most common 30,000 tokens to remove uncommon occurrences that cannot be generalised. Following these steps, queries and answers are then denoted in the dataset with markup tags \(<Q> ... </Q>\) and \(<A> ... </A>\), which are useful for several purposes: (i) to condition the model on separate types of text, (ii) to present the model with queries,
and (iii) to aid in the logic of ending the prediction loop when an answer has been generated.

With regards to the preprocessed data, a batch search of model hyperparameters were implemented for the generative transformer model. Starting from a random weight distribution, topologies of $\{2, 4, 8, 16\}$ attention heads was engineered and attached to one layer of $\{64, 128, 256, 512\}$ rectified linear units. Shallow networks are produced due to the data requirements of deeper models; although alarge dataset was collected, it is relatively close to the minimum requirements of a model following this learning method. In future, given more data, deeper networks could be explored. Models are trained and compared based on the validation metrics of accuracy and loss, with consideration also given to top-$k$ accuracy where $k = 5$ and $k = 10$. Top-$k$ metrics are important for deeper comparison of similarly-performing models, since it is a further measure of how incorrect a wrong prediction is. For example, two models selecting the correct token half of the time will both score 50% accuracy, but one model’s second choice may more often be correct, suggesting that it is on a better track to generalise the data compared to the other.

To conclude the methodology shown in Fig. 2, a general diagram for the process of interfacing with the chatbot and inferring a response from the input query is shown in Fig. 3.

### Table 1: Loss values for the transformer topology tuning experiments.

<table>
<thead>
<tr>
<th>Dense Neurons</th>
<th>Attention Heads</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>64</td>
<td>0.64</td>
</tr>
<tr>
<td>128</td>
<td>0.65</td>
</tr>
<tr>
<td>256</td>
<td>0.65</td>
</tr>
<tr>
<td>512</td>
<td>0.64</td>
</tr>
</tbody>
</table>

### 4 RESULTS AND OBSERVATIONS

In this section, the observed metrics during the topology engineering for the transformer-based chatbots are presented before exploring some examples of its usage after training.

Table 1 and Table 2 show the loss and accuracy metrics for the 16 individual experiments, respectively. Two equally scoring models outperformed all others, which were eight attention heads succeeded by either 64 or 128 rectified linear units. Both of these models could predict the next token 88.65% of the time. Further to loss and accuracy metrics, Tables 3 and 4 show the top-$k$ accuracy for $k = 5$ and $k = 10$, respectively. Beyond the initial results, these tables show us that the option of using 128 neurons in the layer prior to token prediction gives a slightly higher result. These were 96.49% (against 96.41%) and 97.88% (against 97.82%). The 8-headed,
Figure 3: Diagram of the inference process for the trained chatbot model interface.

Table 2: Accuracy values for the transformer topology tuning experiments.

<table>
<thead>
<tr>
<th>Dense Neurons</th>
<th>Attention Heads</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>84.13</td>
<td>86.23</td>
<td>88.65</td>
<td>79.5</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>83.88</td>
<td>85.95</td>
<td>88.65</td>
<td>74.3</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>83.81</td>
<td>85.46</td>
<td>88.02</td>
<td>69.98</td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>84.01</td>
<td>85.44</td>
<td>68.9</td>
<td>63.86</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Top-5 accuracy values for the transformer topology tuning experiments.

<table>
<thead>
<tr>
<th>Dense Neurons</th>
<th>Attention Heads</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>94.86</td>
<td>95.73</td>
<td>96.41</td>
<td>90.89</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>94.84</td>
<td>95.53</td>
<td><strong>96.49</strong></td>
<td>87.15</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>94.78</td>
<td>95.43</td>
<td>96.34</td>
<td>84.26</td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>94.85</td>
<td>95.31</td>
<td>87.7</td>
<td>78.7</td>
<td>79.79</td>
</tr>
</tbody>
</table>

Table 4: Top-10 accuracy values for the transformer topology tuning experiments.

<table>
<thead>
<tr>
<th>Dense Neurons</th>
<th>Attention Heads</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>96.92</td>
<td>97.64</td>
<td>97.82</td>
<td>93.92</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>96.94</td>
<td>97.36</td>
<td><strong>97.88</strong></td>
<td>90.92</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>96.83</td>
<td>97.29</td>
<td>97.77</td>
<td>88.78</td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>96.95</td>
<td>97.23</td>
<td>88.48</td>
<td>84.12</td>
<td></td>
</tr>
</tbody>
</table>

128-neuron model is therefore selected as the best candidate for further exploration.

To provide examples of interaction with the trained model with queries provided by a human and answers output by the model, the following is an excerpt from a conversation about depression:

- **Q**: what is depression
- **A**: clinical depression is a serious condition that negatively affects how a person thinks, feels, and behaves

- **Q**: what are the symptoms of depression
- **A**: the symptoms of depression can be complex and vary widely between people. If you are depressed, you may feel sad, hopeless, and lose interest in things you used to enjoy

- **Q**: how is depression diagnosed
- **A**: the first step to being diagnosed is to visit a doctor for a medical evaluation. Certain medications and some medical conditions such as thyroid disorder can cause similar symptoms as depression

- **Q**: are certain people more likely to get depressed than others
- **A**: you may be more vulnerable to depression if you have certain personality traits such as low self-esteem or being overly self-critical. This may be because of the genes you inherited from your parents, your early life experiences, or both

- **Q**: can you tell me about depression after childbirth
- **A**: sometimes new mothers, fathers, or partners develop depression after having a baby. This is known as postnatal depression, and it is treated in a similar way to other types of depression with talking therapies and antidepressant medicines

In addition, the following is an excerpt from a conversation about anxiety:

- **Q**: is anxiety normal
- **A**: everyone has feelings of anxiety at some point in their life. For example, you may feel worries and anxious about sitting an exam or having a medical test or job interview. During times like these, feeling anxious can be perfectly normal

- **Q**: if I feel too anxious should I see a doctor
- **A**: although feelings of anxiety at certain times are completely normal, you should see a gp if anxiety is affecting your daily life or causing you distress

- **Q**: what are the symptoms of anxiety
- **A**: gad can cause both psychological mental and physical symptoms. These vary from person to person but can include feeling restless or worried, having trouble concentrating, or sleeping dizziness or heart palpitations

- **Q**: what does gad mean

1Note: all grammar has been removed from the training data to enable generalisation of tokens with identical meaning
The use of attention mechanism techniques to build chatbots from an added data preprocessing step. More informative if abbreviations are replaced with definitions as the model can be affected when the majority of terms are abbreviations. Reducing the number of unique tokens via removing grammar aids in training with a dataset of this given size, but results in none being output. In future, more natural conversation would be enabled through either learning from a grammatically-correct dataset, or correcting the chatbot output prior to the response being printed to an interface.

5 CONCLUSION AND FUTURE WORK

In this work, the engineering and applications of transformer-based chatbots are explored to answer questions with a focus on mental health support. Specifically, the focus is on queries surrounding depression and anxiety from respected and verified sources. To conclude this work, chatbots have the potential to play a significant role in supporting people suffering from mental health stigma. The use of attention mechanism techniques to build chatbots from transformers, which are large language models, seem to lead to the creation of engaging conversational systems. The results of this study demonstrate the potential of chatbots to provide easily accessible and anonymous support to people who may otherwise be discouraged from seeking help due to stigma. However, with these findings considered, it is also important to acknowledge the limitations and challenges of using chatbots for mental health support.

More research from medical and psychological backgrounds is needed to fully understand the limitations of chatbots and ensure that they are developed and deployed ethically and responsibly.

Alongside future work regarding ethics, there are also limitations to this study that should be explored. Firstly, data availability is a concern; although we collected a large dataset for this study, this laborious process led to only the minimal amount of data to train such models. In the future, more data could be collected and experiments could be reimplemented to further generalisation. Additionally, methods such as transfer learning and data augmentation could be explored as alternatives to alleviate this limitation. To engineer the topologies, we performed a batch search; this could be further improved through metaheuristic hyperparameter optimisation to automate this process. Although this would likely lead to a better model, it would require far more computational resources and time.

In addition to future experiments, examples such as the chatbot outputting "GAD" (instead of General Anxiety Disorder) show how the model can be affected when the majority of terms are abbreviated within the training data. In the future, the application model may be more informative if abbreviations are replaced with definitions as an added data preprocessing step.

Finally, in conclusion, this study highlights the importance of continuing research and development in the field of mental health technology. By exploring the potential of chatbots to provide support to individuals experiencing depression and anxiety, we can work toward creating innovative and effective solutions to promote mental well-being.

REFERENCES


